IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BIG DATA LEARNING AND DISCOVERY

Received October 17, 2018, accepted October 31, 2018, date of publication November 9, 2018,
date of current version December 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2880233

Disappearing Link Prediction in Scientific Collaboration Networks

BO XU, LU LI, JIAYING LIU, LIANGTIAN WAN, (Member, IEEE), XIANGJIE KONG, (Senior Member, IEEE), AND FENG XIA, (Senior Member, IEEE)

Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China

Corresponding author: Liangtian Wan (wan.liangtian.2015@ieee.org)

This work was supported in part by the Natural Science Foundation of Liaoning Province, China under Grant 201602154, in part by the Fundamental Research Funds for the Central Universities under Grant DUT17RC(3)029, and in part by the Dalian Science and Technology Innovation Fund under Grant 2018J12GX048.

ABSTRACT It is a common sense that both the formation and dissolution of links are the fundamental processes of link dynamics in network organization. Previous studies have analyzed the formation of links with predicting missing links in current networks and new links in the future. However, little attention has been paid to the disappearing link prediction problem. In this paper, we investigate the disappearing link prediction problem. First, we define the disappearing link prediction in scientific collaboration networks. In contrary to the missing link prediction, we use structural similarity indices to estimate the disappearing links through dissimilarity of the node pairs. Then, we propose a novel method called modified preferential attachment (MPA) for predicting disappearing links. MPA is designed based on the preferential attachment considering both links' weights and the different impacts of the nodes' neighbor links. Finally, we evaluate the performance of MPA based on three real scientific collaboration networks extracted from Digital Bibliography & Library Project and American Physical Society datasets. Meanwhile, we explore the performance of the classical similarity methods on disappearing link prediction. The experiment results show that MPA achieves better performance than other classical similarity indices, which verifies the effectiveness of MPA.

INDEX TERMS Disappearing link prediction, scientific collaboration networks, structural similarity.

I. INTRODUCTION

Mining the law of evolution in the network has always been an important subject in the field of network science. Meanwhile, link prediction is an important method to study the evolution of complex networks [1]. It is a commonplace fact that the formation and dissolution of links are two fundamental processes of network dynamic evolution [2]. To the best of our knowledge, most link prediction works focus on link formation by predicting missing links in current networks and new links in the future, and identifying spurious links in networks [3]–[6]. However, only little attention is paid to predict the disappearing links in the future. Fig. 1 shows the difference between missing link prediction and disappearing link prediction. Missing link prediction aims to predict the hidden links in the current network, and disappearing link prediction is to predict disappearing or dissolution links in the future networks. Different from link prediction, disappearing link prediction can help scholars to understand the mechanism of complex network evolution more deeply from another point of view.

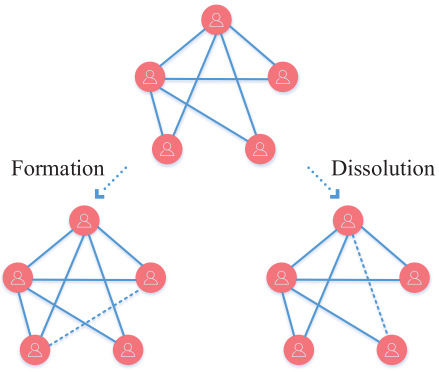


FIGURE 1. The difference between missing link prediction and disappearing link prediction. The left arrow points to the missing link prediction, and the right arrow indicates the disappearing link prediction.

Link disappearance is one of the basic processes of networks evolution. It can represent the ties dissolution which is more rational than ties formation. Comparing with ties

formation which is a simple and quick process, ties dissolution is reasonable and rely more on the previous interactions between targets. The disappearance of links is a universal phenomenon in the real world networks. For example, in scientific collaboration networks, scholar A may stop collaborating with another scholar B [7]. In email networks, the termination of online social behavior (the exchange of emails) reflects the disappearance of links [8]. In online social networks, the phenomenon of disappearance of links has occurred, which is called unfollowing, and researchers have studied this phenomenon in Twitter, Facebook and other online social networks [9]–[13]. Understanding the link disappearance help users know more about how individuals' personal attributes affect the network evolution and how ties be transformed over time. Therefore, the study of disappearing link prediction is of great significance and practical application value.

However, there are few studies focus on disappearing link prediction problem. The current researches mainly focus on link prediction and the analysis of the social phenomena about the relationship dissolution. For link prediction problem, many efforts have been made in this field [14]–[20]. Recently, the issue of structural properties of evolving temporal networks has received considerable critical attention [21]–[23]. Shang *et al.* [24] introduce the temporal prediction accuracy measures and define the evolving link prediction problems. Soares and PrudêNcio [25] propose proximity measures for link prediction based on temporal events. Ozcan and Oguducu [26] propose a novel method for link prediction in dynamic heterogeneous networks based on NARX neural networks.

The approaches for link prediction can be divided into three categories: the similarity-based methods, the likelihood estimation-based methods, and the machine learning-based methods. The similarity-based algorithms are the simplest link prediction methods. They compute the similarity between node pairs by various similarity methods (i.e., Euclidean distance, cosine similarity) and provide an ordered list to predict the likelihood of a link between two nodes. Node similarity methods are one type of the similarity-based methods. They mainly utilize the nodes' attributes and characteristics to describe their personal interests and social behaviors. They usually use similarity methods such as Euclidean distance and cosine similarity to calculate the similarity between two nodes [27], [28]. However, the essential attributes of nodes such as personal information in the social networks are generally implicit and difficult to obtain. Another type of similarity-based methods called structural similarity indices can obtain the similarity between two nodes from the network structure. It is useful when the attributes of nodes are not obtainable [29], [30]. The link prediction method based on likelihood estimation is a more complex type of link prediction methods, which plays an important role in understanding the structural characteristics of the network. They usually use a hierarchical structure model based on likelihood estimation to solve the link prediction problem. However, their computational complexity is relatively high. Therefore, this type of link prediction algorithms cannot be applied to some large-scale networks in the real world. Machine learning-based methods include feature-based classification, probabilistic model, matrix factorization, and so on. They regard the link prediction problem as a supervised classification problem in machine learning. Specially, they mark the relationships between nodes in the network before classification, and then select the appropriate features as the input of the classification algorithm to learn and classify [3], [31]–[37].

While some researchers have examined the analysis of the social phenomena about the relationship dissolution, the studies only stay in the discovery and analysis of the reason for links dissolution. For example, in online social networks, the behavior that users stops the online social relationship with another user is known as ''unfollowing''. Some works begin to research in this area. Shang *et al.* [38] define four types of users and interactions including disappear nodes in the dynamic social network. Kwak *et al.* [9] analyze the dynamics of the unfollowing behavior in Twitter to understand the online relationship dissolution. They find that several factors, including the reciprocity of the relationship, the duration of the relationships, the followees' informativeness, and the overlap of relationships, are the key factors on unfollowing. They further build a logistic regression model considering two different sets of factors, including structure and action [10]. After a quantitative analysis, they find that people enjoy getting more attention than giving. Kivran-Swaine *et al.* [12] explore the influence of network structure alone on link dissolution in Twitter. They use multilevel logistic regression to study the impact of the network properties of the seed, followers, and dyad on the breaking of links. Then, they contact several sociology concepts such as link strength, embeddedness, and status to analyze the results of the unfollowing behavior. Xu *et al.* [11] study the differences of relational and informational factors on unfollowing relations in Twitter. They use an actor-oriented model (SIENA) to explore the impacts of reciprocity, social status, embeddedness, topic-homophily, and informativeness on the link dissolution. Sibona and Walczak [39] analyze surveys conducted online, and find the role of the friend request in unfollowing behavior. Quercia *et al.* [13] apply some factors in sociology such as age, gender, and personality traits to study the unfollowing behavior in Facebook. Their findings are consistent with previous analyses of Twitter.

All of the above works focus on the unfollowing behavior on Twitter, Facebook and other online social networks. The disappearing links in these scenarios may influence the interactions between the users and the evolution of the overall network. However, they mainly analyze the universality of unfollowing and explore the possible causes of this phenomenon. Existing studies only focus on the discovery and analysis of the reason for links dissolution. Few researchers try to analyze the unfollowing behavior from the perspective of the disappearance of links in network science.

In this paper, based on the assumption that the lower of similarity between the node pairs that constitute links, the higher possibility of the disappearance of links, we investigate the disappearing link prediction in collaboration networks. Firstly, we define the disappearing link prediction in scientific collaboration networks, and analyze the process of disappearance of links in the scientific collaboration network constructed by a real dataset according to the definition. Next, we propose a novel disappearing link prediction method called MPA (Modified Preferential Attachment), which combines links' weights and types together to calculate the structural similarity. We divide the neighbor links of the target node pairs into three types based on the different neighbor nodes. We observe that links between the target nodes and common neighbors have negative effects for predicting the possibility of the disappearance of target links. By using the reciprocal function to weaken the impact of common neighbor links, we design the MPA index. Then, we evaluate the performance of the proposed method in three real scientific collaboration networks, which are constructed from Digital Bibliography & Library Project (DBLP) and American Physical Society (APS) datasets. Moreover, we explore the performance of the classical similarity methods on disappearing link prediction, including Common Neighbors index [40], Adamic-Adar index [41], Resource Allocation index [42], Preferential Attachment index [43], Local Path index [44], Katz index [45], and Random Walk with Restart index [46]. The experiment results show that compared with the classical similarity indices, MPA index achieves better prediction performance, and demonstrate the effectiveness of the proposed method in disappearing link prediction. The contributions of this paper can be summarized briefly as follows:

- We define the disappearing link prediction problem in the scientific collaboration networks.
- We find that the links between nodes and their common neighbors have a negative impact on disappearing link prediction.
- We propose a novel disappearing link prediction model, which can explore the disappearance of links in scientific collaboration networks.

The remaining contents of this paper are organized as follows. Section II describes the definition of the disappearing link prediction and gives a case of obtaining disappearing link set from the DBLP dataset. Section III proposes the novel method MPA. Section IV explores the performance of the proposed method in three real scientific collaboration networks by comparing it with seven similarity indices. Finally, Section V concludes the paper.

II. DISAPPEARING LINKS PREDICTION PROBLEM

In this section, we describe the definition of the disappearing link prediction in scientific collaboration networks. Then, we introduce the process of obtaining the disappearing links in scientific collaboration network constructed by the DBLP dataset, and analyze the evolution of the disappearance of links in the network.

A. PROBLEM DEFINITION

We first consider the scientific collaboration network as an undirected network $G(V, E)$ at a particular time slice $[t_0, t]$ $(t_0 < t)$, where V and E are sets of nodes and links, respectively. In this network, the nodes represent the authors in the dataset and an edge between authors means that they have co-authored a paper. The link connecting nodes A and B at a particular time t is defined as a disappearing link if this link will not be active again for a long time till future time point t' $(t \ll t')$.

In other words, a disappearing link means the end of the connected relationship between two nodes. In particular, a disappearing link in scientific collaboration networks means the termination of a collaborative relationship between scholars. Those links that disappeared briefly and then reconnected are not the disappearing links as defined in this paper. Therefore, the disappearing link prediction aims to predict the probability of the disappearance of a link within a sufficiently long time. For missing link prediction problem, the greater the similarity between node pairs, the greater the likelihood of the link generation. Conversely, for disappearing link prediction, we assume that the higher the dissimilarity of node pairs, the higher the possibility of disappearance of links.

In disappearing link prediction, the observed links are divided into two parts: the links in the disappearing set E^D are never connected again, while the remaining links in the set E^R are the existent links.

B. DISAPPEARING LINKS IN SCIENTIFIC COLLABORATION NETWORKS

According to the definition of disappearing link, we can get the disappearing link set from scientific collaboration networks. In this section, we use DBLP dataset to construct the scientific collaboration network [47]. Meanwhile, we analyze the phenomenon of the disappearance of links in the network. The DBLP dataset describes the collaborative relationship between the authors in the computer science discipline. If author A and author B co-author a paper, there is a corresponding record in the dataset. Each record contains the authors' ID and the publication date of the co-authored paper. According to some records of this dataset, we can construct a scientific collaboration network. In this network, authors are the nodes. It should be noted that the weights of links may be greater than 1, which indicates that two authors co-authored several papers.

In disappearing link prediction problem, we only pay attention to the disappearance of links in the network, so it is more reasonable to choose the connected network for the study of disappearing link prediction. In this paper, the maximal connected subgraph, which is the largest connected component of the scientific collaboration network is selected as the object of the study.

Taking the period from 1990 to 1992 as an example, the undirected scientific collaboration network G_c is constructed by extracting the collaborative relationship data from

the DBLP dataset in this period. After the choice of the time, the remaining time is long enough to explore the disappearing link prediction problem. The size of G_c is large, and the number of nodes and edges are 64,317 and 104,644, respectively.

Then, we extract the maximal connected subgraph (MCS) of this scientific collaboration network G_c. In the MCS, the number of nodes is 26,336, and the number of edges is 57,174.

Next, we crawl the co-authorship information of scholars in the MCS over the next 20 years. By using the time information of the dataset, we separate the annual co-author relationships in the MCS network and build 20 scientific collaboration networks. According to the definition of the disappearing link, we calculate the ratio between the number of disappearing links in each year and the total number of links of MCS, and the results are shown in Fig. 2.

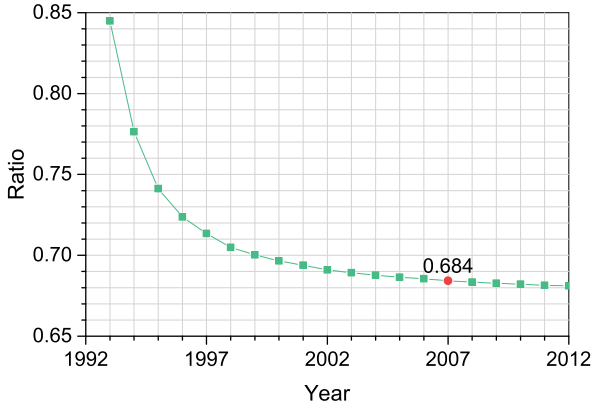


FIGURE 2. The ratio of disappearing links in each year over the next 20 years.

Because the MCS network is connected, the ratio of the disappearing link in 1992 is 0. As shown in Fig. 2, in 1993, the ratio of broken links is about 84.5%, which indicates the universality of disappearance of links. But some broken links may be connected again in the next year or other future time, after which time we do not classify this type of link as the disappearing links. So with the change of time, the link between nodes in the network is unstable. In this network, we only need to find out the links that never collaborate again with each other from 1993. We find that after the year 2007, the change of the ratio is less than 0.001 in each year, and the ratio almost remains at 68.4%. This means that 15 years after the formation of the MCS network, few nodes of broken links will collaborate again. In other words, the remaining broken links are the disappearing links, and we can choose anytime after 15 years to obtain the disappearing links. Here, we choose disappearing links at the year 2007 as the real disappearing links. Thus, we get the disappearing link set for disappearing link prediction experiments.

The evolution of the MCS scientific collaboration network is shown in Fig. 3. As shown in Fig. 3(a), the MCS is extracted from the scientific collaboration network from 1990 to 1992. Any pair of nodes in this figure has at least a path from one node to another. In Fig. 3(b), the network produces a large number of broken edges in 1993. Evolving with time, some broken edges are connected again. As long as the broken edges produce at least once collaboration, then the broken edges are not the disappearing links. And we mark these broken edges that can be reconnected as the connected edges. Therefore, the broken edges always decrease with time. In Fig. 3(c), until 2007, the broken links is almost fixed, and these broken edges are named disappearing links in the MCS network.

III. METHOD

In this section, we first analyze the preferential attachment indices in this work. Then, we give the overall description of our model.

A. PREFERENTIAL ATTACHMENT INDEX ANALYSIS

In this study, the hypothesis of disappearing link prediction is that the lower the similarity of node pairs, the higher the possibility of disappearance of links. Inspired by Preferential Attachment index (PA) [43] in link prediction problem, we consider that different neighbor links of node pairs and their weight may have different effects on target links in the disappearing link prediction, so we design a new index MPA (Modified Preferential Attachment), which is suitable for the disappearing link prediction. The PA index indicates that the probability of a new link is proportional to degrees of node pairs. The PA_w index (hereinafter termed PA model) is the weighted form of PA index. The PA model is expressed as

$$s_{xy} = s_x s_y = \sum_{i\in\Gamma(x)} w_{ix} \times \sum_{j\in\Gamma(y)} w_{jy}. \tag{1}$$

From Equation (1), it can be found that the PA model overlooks the effect of different link types between node pairs. In reality, different types of neighbor links may have different influences for the target links in disappearing link prediction. Fig. 4 briefly shows the ego network of node pair x and y. In this figure, there are three types of links in target nodes' ego networks.

Thus, we divide the links of target nodes into three types as shown in Fig. 4, including:

1) L_1: the link connecting node x and node y;
2) L_2: the links connecting node x (or y) and its common neighbors with node y (or x);
3) L_3: the other neighbor links.

We try to explore the effect of the ego network structure on disappearing link prediction by calculating the PA similarity with different link types. Meanwhile, the link weight w is considered as well. We add three parameters λ, μ, and ν on L_1, L_2, and L_3, and the modified s_x and s_y can be calculated as

$$\begin{aligned} s_x &= s_x^m = \lambda L_1 + \mu L_2 + \nu L_3 \\ &= \lambda w_{xy} + \mu \sum_{i\in\Gamma(x)\cap\Gamma(y)} w_{ix} + \nu \sum_{j\in\{\Gamma(x)-\Gamma(x)\cap\Gamma(y)-y\}} w_{jx}, \end{aligned} \tag{2}$$

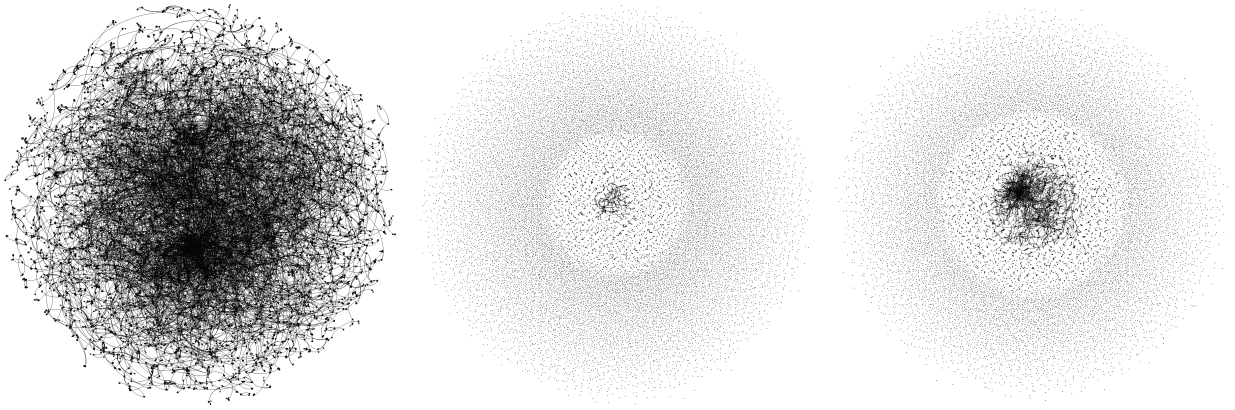


FIGURE 3. The evolution of the MCS scientific collaboration network. (a) 1990 - 1992. (b) 1993. (c) 2007.

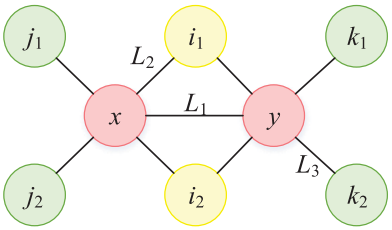


FIGURE 4. The example of link type classification.

and

$$s_y = s_y^m = \lambda L_1 + \mu L_2 + \nu L_3 = \lambda w_{xy} + \mu \sum_{i \in \Gamma(x) \cap \Gamma(y)} w_{iy} + \nu \sum_{k \in \{\Gamma(y) - \Gamma(x) \cap \Gamma(y) - x\}} w_{ky}, \tag{3}$$

respectively, where parameters λ, μ, and ν are either 1 or 0. They are used to distinguish whether the direct link between the target node pairs, the link between the target node and the common neighbor, and the link between the target node and the other neighbors contributes to the disappearing link prediction problem. Thus, we have seven strategies (A1 to A7 in Fig. 5) to calculate the PA similarity on scientific collaboration network constructed by DBLP dataset from 1990 to 1992, and get the precisions. The number of predicted links is set to 10,000. Each three-tuple under A1 to A7 in Fig. 5 corresponds to the three parameters λ, μ, and ν that control the existence of L_1, L_2 and L_3. For example, A1(1,0,0) indicates a strategy that contains only L_1.

We find that common neighbor links L_2 lead to the lowest precision in disappearing prediction links (Strategy A2). Strategies containing L_2 (A2, A4, A6, and A7) will lead to worse results. Meanwhile, Strategy A5 without L_2 achieves the best performance. So it can be known that the weight of common neighbor links has a negative effect on disappearing link prediction.

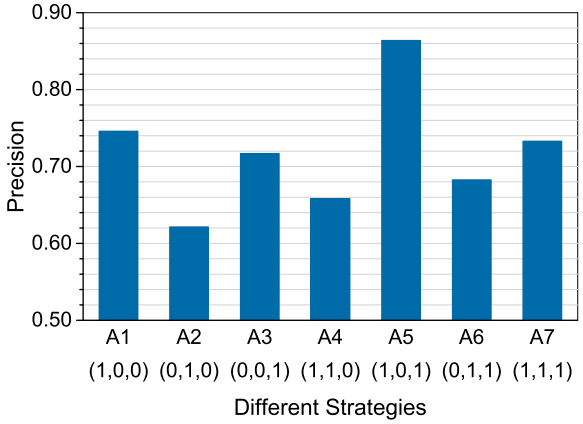


FIGURE 5. The precisions of PA similarity with seven combinations of L_1, L_2 and L_3. The value of three-tuple is corresponding to the three parameters λ, μ and ν, respectively.

B. MODEL CONSTRUCTION

Thus, our MPA model is proposed to weaken the impact of L_2 and improve the effective of disappearing link prediction. Many ways can weak the negative impact of L_2, in this study we use its reciprocal function $1/L_2$ to deal with this problem. The MPA model is defined as follows:

$$\mathrm{MPA}_{xy} = \prod_{\eta \in \{x,y\}} s_\eta = \prod_{\eta \in \{x,y\}} \{ \sum_{i \in \Gamma(x) \cap \Gamma(y)} \frac{1}{w_{i\eta}} + \sum_{k \in \{\Gamma\eta - \Gamma(x) \cap \Gamma(y)\}} w_{k\eta} \}. \tag{4}$$

It is worth mentioning that when nodes x and y have no common neighbors, $1/w_{i\eta}$ is set as 1 in order to avoid the meaningless results.

We can get the similarity between two target nodes based on the MPA model. We rank all node pairs by MPA in an inverse order. Based on the assumption that less similarity of node pairs will lead to a higher possibility of the disappearance of links, the node pair with the smallest similarity

is mostly like to break. The top L links are predicted as disappearing links. When the number of predicted links is set to 10,000, the precision of MPA is 0.873, and it is bigger than that of the best Strategy A5 (0.864). So the design of MPA is effective and reasonable.

IV. EXPERIMENTS

To evaluate the effectiveness of the proposed model on the task of disappearing links prediction, we conduct a series of experiments. In this section, we will present the details about how we carry out the experiments to certify the accuracy and the effectiveness of our model. All the experiments are conducted 128G serve with the Python 2.7 simulation environment.

A. DATASETS

In order to verify the effectiveness of the proposed method, we apply the MPA model to predict the disappearing links in the scientific collaboration network of Digital Bibliography & Library Project (DBLP). In addition, we apply the MPA model in American Physical Society (APS) datasets as well. We choose two journals which are Physical Review A (PRA) and Physical Review B (PRB) to validate the universality of MPA.

B. DATASETS PRE-PROCESSING

First, we extract the collaboration information at a particular time slice $[t_0, t]$ to construct the undirected collaboration network G_c. Since we aim to study the problem of disappearing link prediction, we only focus on the disappearance of links in the network. We select the connected networks of which there is a path between any two nodes for experiments. We extract the maximum connect component of G_c to conduct the experiments.

We extract co-authorship information of scholars in DBLP, PRA, and PRB from two time intervals, which are from 1987 to 1989 and from 1990 to 1992. The preprocessing of PRA and PRB is similar with the DBLP dataset. The basic information of the MCS networks in DBLP, PRA, and PRB are shown in TABLE 1. We find that the disappearing link ratios of two scientific collaboration networks constructed by PRA and PRB remain steady after 15 years as well. The ratio in TABLE 1 represents the disappearing link ratios after 15 years.

TABLE 1. The Basic Features of The MCS Networks in DBLP, Physical Review A, and Physical Review B.

Dataset	Year	$\|V\|$	$\|E\|$	Ratio
DBLP		13,397	27,759	68.9%
PRA	1987-1989	1,904	5,710	78.2%
PRB		9,485	33,495	79.5%
DBLP		26,336	57,174	68.4%
PRA	1990-1992	1,991	6,387	80.1%
PRB		11,343	40,094	78.2%

C. EVALUATION METRICS

The common metrics used to quantify the accuracy of prediction algorithms are AUC (area under the receiver operating characteristic curve) [48] and precision [49].

The definitions of AUC and precision in disappearing link prediction are different from that defined in missing link prediction. The AUC value can be understood as the probability that a randomly chosen link in E^D (i.e., a disappearing link) is given a lower score than a randomly chosen link in E^R. To obtain the value of the AUC, we randomly pick a disappearing link and an existent link in the observed network, and compare their scores. If among all possible comparisons m, there are m' times the disappearing link having a lower score than the existent link and m'' times the disappearing link and the existent link having the same score, the AUC value is written as follows

$$\text{AUC} = \frac{m' + 0.5m''}{m}. \tag{5}$$

Obviously, if all the link scores are randomly generated, the AUC value would be about 0.5. Therefore, the degree of the AUC value of the algorithm over 0.5 indicates that the algorithm performs better than the random selection methods.

The calculation of precision also has a small change, which is sorting the similarity from small to large. If we pick up the top-L links (i.e., the number of predicted links), among which L_d links are disappearing links, then the precision is expressed as

$$\text{Precision} = \frac{L_d}{L}. \tag{6}$$

Clearly, the size of this value is related to parameter L. For a given L, higher precision indicates higher prediction accuracy.

We also use the precision-recall curves to reflect the performance of the proposed model. In our experiments, we calculate recall as:

$$\text{Recall} = \frac{L_d}{L_{dis}}. \tag{7}$$

where L_{dis} is the total number of disappearing links.

D. RESULTS AND DISCUSSIONS

For disappearing link prediction problem, the hypothesis of this study is that the lower the similarity of node pairs, the higher the possibility of disappearance of links.

1) BASELINE METHODS

In order to evaluation the performance of MPA, we choose seven classical structural similarity indices as baselines:

1) Common Neighbors (CN) [40] This index is defined as the counting of common neighbors.
2) Adamic-Adar (AA) [41] The idea of AA index is that the contribution of the low-degree common neighbors is bigger than that of the high-degree common neighbors.

TABLE 2. The similarity indices.

Index	Definition
CN	$s_{xy} = \lvert\Gamma(x) \cap \Gamma(y)\rvert$
AA	$s_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log k_z}$
RA	$s_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log k_z}$
PA	$s_{xy} = k_x \times k_y$
LP	$S = A^2 + \alpha \cdot A^3$
Katz	$S = (I - \beta \cdot A)^{-1} - I$
RWR	$s_{xy} = q_{xy} + q_{yx}$, $\vec{q}_x = (1-c)(I - cP^T)^{-1}\vec{e}_x$, $P_{xy} = \frac{a_x}{k_x}$
CN_w	$s_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w_{xz}+w_{zy}}{2}$
AA_w	$s_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w_{xz}+w_{zy}}{2\log(1+s_z)}$
RA_w	$s_{xy} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{w_{xz}+w_{zy}}{2s_z}$
PA_w	$s_{xy} = \sum_{i\in\Gamma(x)} w_{ix} \times \sum_{j\in\Gamma(y)} w_{jy}$
LP_w	$S = W^2 + \alpha \cdot W^3$
Katz_w	$S = (I - \beta \cdot W)^{-1} - I$
RWR_w	$P_{xy} = \frac{w_{xy}}{s_x}$

3) Resource Allocation (RA) [42] This index assumes that each common neighbor has a unit resource, and will assign the resource on average to all its neighbors.
4) Preferential Attachment (PA) [43] This index defines that the similarity of two nodes is proportional to the product of their respective degrees.
5) Local Path (LP) [44] This index takes into account the local paths with length 2 and length 3.
6) Katz [45] This index is a global metric, and it sums over all the paths between two nodes.
7) Random Walk with Restart (RWR) [46] This index is a random walk model, considering the case of returning the starting node.

TABLE 2 summarizes the unweighted and weighted formulas of above seven similarity indices in detail. $\Gamma(x)$ is the set of neighbors of node x, and k_z is the degree of node z. In LP index, α is a free parameter that controls the weights of the paths with length 3, and A is the adjacency matrix. In Katz index, β is a free parameter that controls the weights of all the paths, and I is the identity matrix. In RWR index, P denotes the transition matrix. The element $P_{xy} = 1/k_x$ if x and y are connected, otherwise $P_{xy} = 0$. s_z denotes the strength of the node z, i.e., the sum of the weights of all links connected to the node z, and w_{xz} represents the weight of the link which is connected by nodes x and z. W is the adjacency matrix in weighted form. In other words, the elements of the adjacency matrix represent the weights of the links. For RWR index, the transition matrix is modified as the weighted form, as shown in TABLE 2, and the remaining calculation process is consistent with its unweighted form.

2) COMPARISON ON AUC

The prediction accuracies on three datasets measured by AUC are shown in TABLE 3. The parameters $\alpha = 10^{-3}$ for LP index, $\beta = 10^{-3}$ for Katz index, and $c = 0.85$ for RWR index.

As shown in TABLE 3, the AUC values of the MPA model in tree datasets are higher than 0.5 in both unweighted and weighted networks whenever from 1987 to 1989 and from 1990 to 1992. In unweighted networks, except PRB network from 1987 to 1989, MPA model achieves the biggest AUC score. In PRB network from 1987 to 1989, the AUC value of MPA is second only to RA index. In weighted networks, MPA model is the first or second only to Katz index. Combining the results, we can find that MPA index has a better prediction effect.

Then, we analyze AUC scores on each dataset. For DBLP, the AUCs of MPA model are significantly higher than all the other seven indices in both unweighted and weighted networks, which indicates that MPA has the best performance in the DBLP dataset. In the PRA network from 1990 to 1992, the Katz index is the best index in the weighted network. MPA is only slightly smaller than Katz. Overall, MPA achieves good performance in PRA datasets. In the PRB network from

TABLE 3. The prediction accuracy measured by AUC.

Indices	DBLP		PRA		PRB	
	1987-1989	1990-1992	1987-1989	1990-1992	1987-1989	1990-1992
CN	0.4298	0.4277	0.4877	0.4927	0.5066	0.5103
AA	0.4372	0.4337	0.4984	0.5112	0.5346	0.5292
RA	0.4451	0.4414	0.4969	0.5473	**0.5604**	0.5491
PA	0.4990	0.5108	0.5391	0.5282	0.5240	0.5409
LP	0.4378	0.4385	0.4921	0.4945	0.5068	0.5123
Katz	0.4376	0.4383	0.4921	0.4944	0.5067	0.5123
RWR	0.4927	0.4776	0.4739	0.5064	0.5094	0.4919
MPA	**0.6072**	**0.6231**	**0.5985**	**0.5480**	0.5411	**0.5554**
CN_w	0.4596	0.4750	0.5160	0.5214	0.5336	0.5412
AA_w	0.4597	0.4686	0.5164	0.5668	0.5818	0.5694
RA_w	0.4585	0.4584	0.5177	0.5321	0.5493	0.5516
PA_w	0.5457	0.5700	0.5783	0.5609	0.5556	0.5721
LP_w	0.4916	0.5154	0.5422	0.5463	0.5543	0.5643
Katz_w	0.5730	0.5995	0.6023	**0.6024**	**0.6146**	**0.6149**
RWR_w	0.5388	0.5320	0.5148	0.5452	0.5577	0.5379
MPA_w	**0.6437**	**0.6634**	**0.6509**	0.5981	0.5821	0.5822

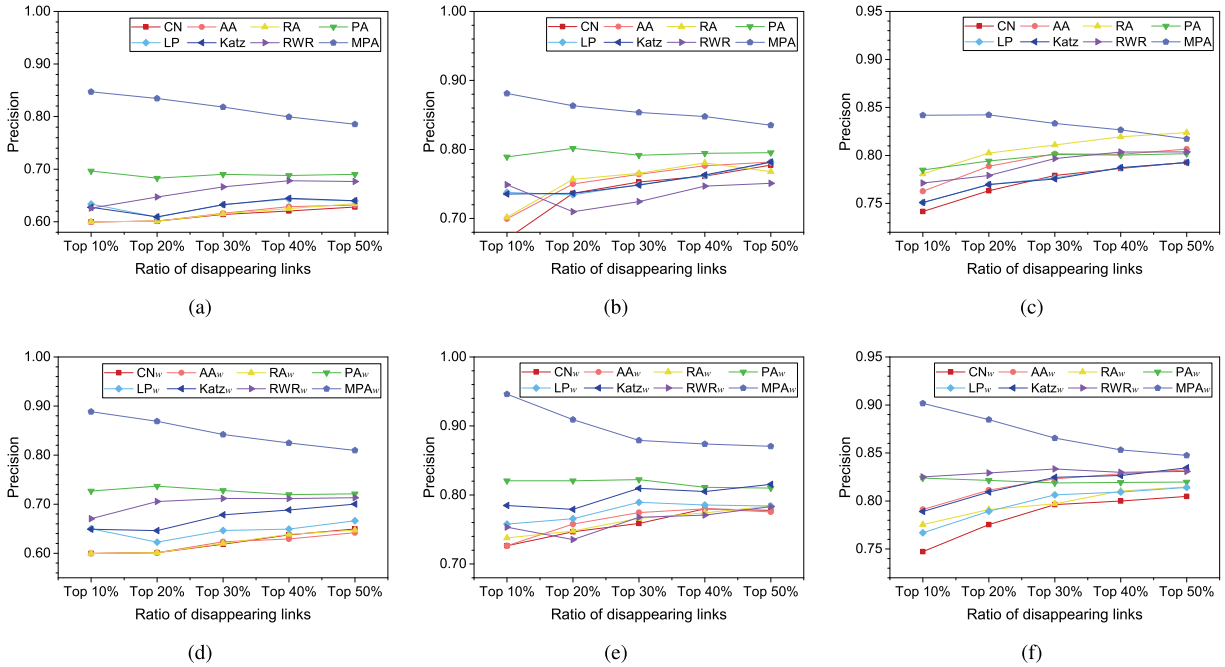


FIGURE 6. The prediction accuracy measured by precision from 1987 to 1989. (a) Unweighted network in DBLP. (b) Unweighted network in PRA. (c) Unweighted network in PRB. (d) Weighted network in DBLP. (e) Weighted network in PRA. (f) Weighted network in PRB.

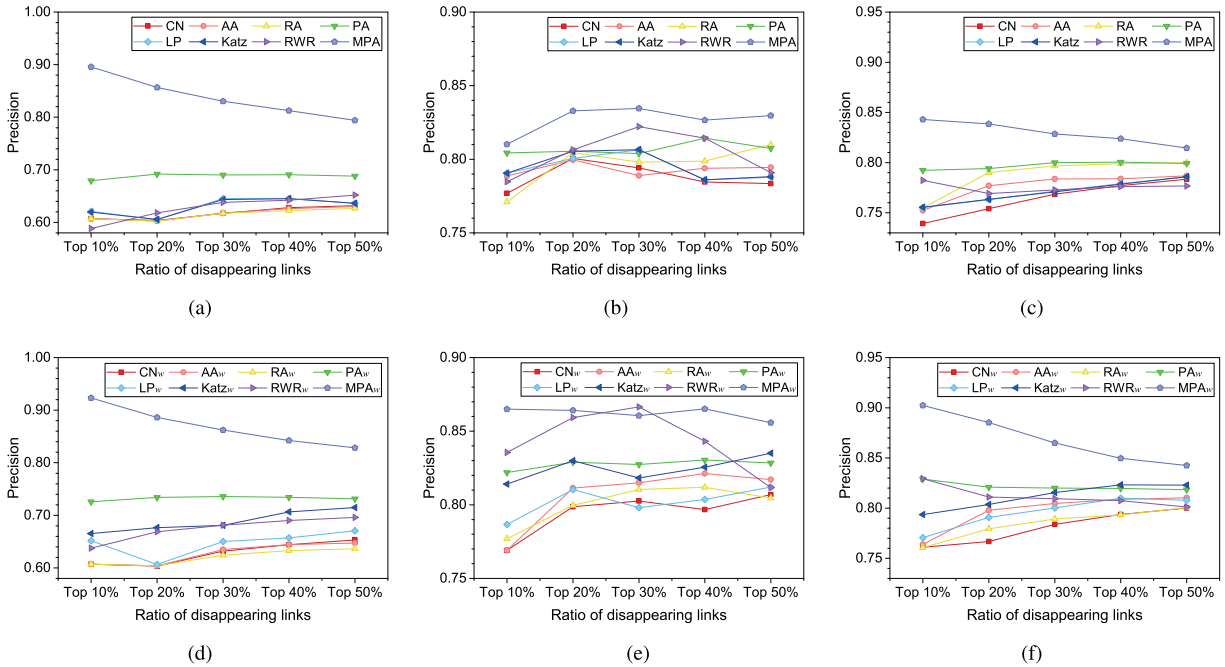


FIGURE 7. The prediction accuracy measured by precision from 1990 to 1992. (a) Unweighted network in DBLP. (b) Unweighted network in PRA. (c) Unweighted network in PRB. (d) Weighted network in DBLP. (e) Weighted network in PRA. (f) Weighted network in PRB.

1987 to 1989, the RA index is the best one in the unweighted network. But the AUC value of MPA is significantly bigger than RA in weighted network. In the weighted network, the AUC of Katz index is highest, and MPA is the second one. But the AUC of MPA is significantly higher than Katz in the unweighted network. In the PRB network from 1990 to 1992, MPA is only slightly smaller than Katz in the weighted network. Therefore, in the PRB dataset, MPA has better performance. In summary, these results show that MPA model is effective in the task of disappearing link prediction. It has the best or nearly the best prediction effect comparing with the other seven indices.

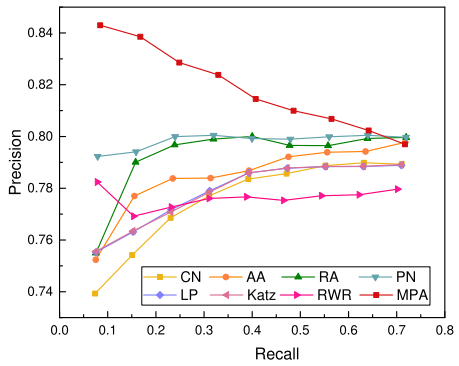
3) COMPARISON ON PRECISION

The number of predicted links mentioned in Equation (6) is set as $k\%$ ($k = 10, 20, 30, 40, 50$), where $k\%$ means the proportion of all the disappearing links in MCS scientific collaboration networks. The accuracies of three datasets measured by precision are shown in Fig. 6 and Fig. 7.

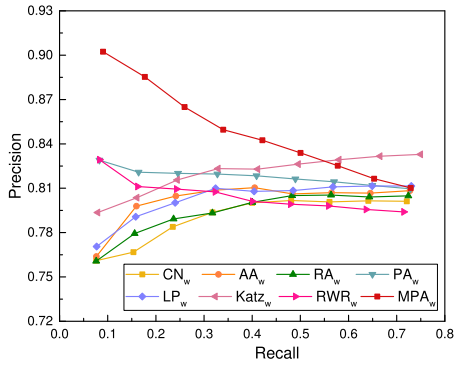
Fig. 6 shows the prediction accuracy measured by precision from 1987 to 1989. Compared with the unweighted network, the precisions of all indices improve in weighted networks. MPA achieves the best performance when predict the top 10% disappearing links, and the precision gradually drops with the increasing ratio of disappearing links. Overall, MPA performs better than the other seven indices. From Fig. 6(c) we can see that, MPA achieves better performance than the other seven indices except slightly smaller than RA at the top 50%. Furthermore, it has the best precisions in the PRB dataset.

As shown in Fig. 7, MPA also has best performance from 1990 to 1992. Fig. 7(e) shows that MPA remains steady from top 10% to top 50%, and it is better than the other seven indices except slightly smaller than RWR at top 30%. But RWR index drops sharply after top 30%. As a whole, the MPA has the best precisions in PRA dataset. The performance of MPA in other situations are best.

Overall, experimental results indicate that MPA has better prediction performance in both unweighted and weighted networks, and MPA is applicable and effective in the disappearing link prediction problem.



(a)



(b)

FIGURE 8. The precision-recall curves for different measures in the PRB dataset from 1990 to 1992. (a) Unweighted network in PRB. (b) Weighted network in PRB.

4) COMPARISON ON PRECISION-RECALL CURVES

Fig. 8 shows the precision-recall curves for different measures in the PRB dataset from 1990 to 1992. Note that, in the experiments, the number of predicted links mentioned in Equation (6) is set as $k\%$ ($k = 10$, 20, 30, 40, 50, 60, 70, 80, 90), where $k\%$ means the proportion of all the disappearing links in the PRB networks. The precision-recall curves for the unweighted PRB network and the weighted PRB network are shown in Fig. 8(a) and Fig. 8(b), respectively.

From the results we can see that MPA has the best prediction performance in both unweighted and weighted networks comparing with other seven measures. It indicates that the task of disappearing link prediction can achieve significant and stable improvement by taking advantages of our proposed method MPA.

V. CONCLUSION

In this paper, we investigate the disappearing link prediction problem in scientific collaboration networks. Firstly, we define the disappearing links and disappearing link prediction in scientific collaboration networks. According to the definition of the disappearing links, we introduce the process of obtaining the disappearing link set from a scientific collaboration dataset. Then, we propose a novel disappearing link prediction method, MPA, which consider the different effects of node pairs' neighbor links. Finally, we evaluate the performance of MPA in disappearing link prediction experiments on three real-world scientific collaboration networks. Experimental results show that MPA index has the best or near best prediction effect compared with seven classical structural similarity indices. Experimental results of MPA and classical similarity indices also verify the rationality of the assumption that the higher the dissimilarity between node pairs, the higher possibility of the disappearance of links. Meanwhile, we explore the potentials of the similarity-based idea for the disappearing link prediction.

In the future, we would like to do more research to further explore more details of the disappearing link prediction. In addition, we will evaluate the applicability of this model on other datasets and other types of networks such as directed networks and heterogeneous networks.

REFERENCES

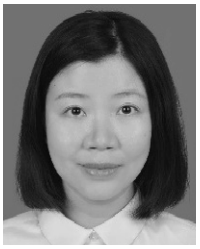
[1] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, 2007.

[2] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: The state-of-the-art," *Sci. China Inf. Sci.*, vol. 58, no. 1, pp. 1–38, Jan. 2015, doi: 10.1007/s11432-014-5237-y.

[3] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 52, pp. 22073–22078, 2009.

[4] L. Pan, T. Zhou, L. Lü, and C.-K. Hu, "Predicting missing links and identifying spurious links via likelihood analysis," *Sci. Rep.*, vol. 6, Mar. 2016, Art. no. 22955.

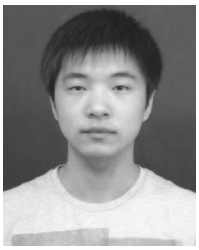
[5] P. Luo, Y. Li, C. Wu, and K. Chen, "Detecting the missing links in social networks based on utility analysis," *J. Comput. Sci.*, vol. 16, pp. 51–58, Sep. 2016.
[6] R. Eyal, A. Rosenfeld, S. Sina, and S. Kraus, "Predicting and identifying missing node information in social networks," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 3, 2014, Art. no. 14.
[7] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars' academic ages," *Scientometrics*, vol. 112, no. 1, pp. 329–343, 2017.
[8] G. Kossinets and D. J. Watts, "Origins of homophily in an evolving social network," *Amer. J. Sociol.*, vol. 115, no. 2, pp. 405–450, 2009.
[9] H. Kwak, H. Chun, and S. Moon, "Fragile online relationship: A first look at unfollow dynamics in Twitter," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, New York, NY, USA, 2011, pp. 1091–1100.
[10] H. Kwak, S. B. Moon, and W. Lee, "More of a receiver than a giver: Why do people unfollow in Twitter?" in *Proc. 6th Int. AAAI Conf. Weblogs Social Media (AAAI)*, 2012, pp. 499–502. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4598
[11] B. Xu, Y. Huang, H. Kwak, and N. Contractor, "Structures of broken ties: Exploring unfollow behavior on Twitter," in *Proc. Conf. Comput. Supported Cooperat. Work (CSCW)*, New York, NY, USA, 2013, pp. 871–876.
[12] F. Kivran-Swaine, P. Govindan, and M. Naaman, "The impact of network structure on breaking ties in online social networks: Unfollowing on Twitter," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, New York, NY, USA, 2011, pp. 1101–1104.
[13] D. Quercia, M. Bodaghi, and J. Crowcroft, "Loosing 'friends' on Facebook," in *Proc. 4th Annu. ACM Web Sci. Conf. (WebSci)*, New York, NY, USA, 2012, pp. 251–254.
[14] M. Al Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social Network Data Analytics*, C. Aggarwal, Ed. Boston, MA, USA: Springer, 2011.
[15] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2011, pp. 1100–1108, doi: 10.1145/2020408.2020581.
[16] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in *Proc. IEEE Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2011, pp. 121–128.
[17] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "MVCwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 364–375, Sep. 2014.
[18] L. Zhu, D. Guo, J. Yin, G. Ver Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2765–2777, Oct. 2016.
[19] S.-Y. Tan, J. Wu, L. Lü, M.-J. Li, and X. Lu, "Efficient network disintegration under incomplete information: The comic effect of link prediction," *Sci. Rep.*, vol. 6, Mar. 2016, Art. no. 22916.
[20] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, "Community detection, link prediction, and layer interdependence in multilayer networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, no. 4, p. 042317, 2017.
[21] K.-K. Shang, W.-S. Yan, and M. Small, "Evolving networks—Using past structure to predict the future," *Phys. A, Statist. Mech. Appl.*, vol. 455, pp. 120–135, 2016.
[22] P. Holme and J. Saramäki, "Temporal networks," *Phys. Rep.*, vol. 519, no. 3, pp. 97–125, 2012.
[23] T. Murata and S. Moriyasu, "Link prediction of social networks based on weighted proximity measures," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.* Washington, DC, USA: IEEE Computer Society, Nov. 2007, pp. 85–88.
[24] K.-K. Shang, M. Small, X.-K. Xu, and W.-S. Yan, "The role of direct links for link prediction in evolving networks," *Europhys. Lett.*, vol. 117, no. 2, p. 28002, 2017.
[25] P. R. S. Soares and R. B. PrudêNcio, "Proximity measures for link prediction based on temporal events," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6652–6660, 2013.
[26] A. Ozcan and S. G. Oguducu, "Link prediction in evolving heterogeneous networks using the NARX neural networks," *Knowl. Inf. Syst.*, vol. 55, no. 2, pp. 333–360, 2018.
[27] P. Bhattacharyya, A. Garg, and S. F. Wu, "Analysis of user keyword similarity in online social networks," *Social Netw. Anal. Mining*, vol. 1, no. 3, pp. 143–158, 2011.
[28] C. G. Akcora, B. Carminati, and E. Ferrari, "User similarities on social networks," *Social Netw. Anal. Mining*, vol. 3, no. 3, pp. 475–495, 2013.
[29] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011.
[30] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning, "Identifying anomalous citations for objective evaluation of scholarly article impact," *PLoS ONE*, vol. 11, no. 9, p. e0162364, 2016.
[31] N. Benchettara, R. Kanawati, and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks," in *Proc. IEEE Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2010, pp. 326–330.
[32] D. Davis, R. Lichtenwalter, and N. V. Chawla, "Supervised methods for multi-relational link prediction," *Social Netw. Anal. Mining*, vol. 3, no. 2, pp. 127–141, Jun. 2013, doi: 10.1007/s13278-012-0068-6.
[33] M. Fire, L. Tenenboim-Chekina, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust, IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 73–80.
[34] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.
[35] T. Vallès-Català, F. A. Massucci, R. Guimerà, and M. Sales-Pardo, "Multilayer stochastic block models reveal the multilayer structure of complex networks," *Phys. Rev. X*, vol. 6, no. 1, p. 011036, 2016.
[36] A. K. Menon and C. Elkan, "Link prediction via matrix factorization," in *Machine Learning and Knowledge Discovery in Databases*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Berlin, Germany: Springer, 2011.
[37] Q. Yang, E. Dong, and Z. Xie, "Link prediction via nonnegative matrix factorization enhanced by blocks information," in *Proc. IEEE 10th Int. Conf. Natural Comput. (ICNC)*, Aug. 2014, pp. 823–827.
[38] K.-K. Shang, W.-S. Yan, and X.-K. Xu, "Limitation of degree information for analyzing the interaction evolution in online social networks," *Int. J. Mod. Phys. C*, vol. 25, no. 10, p. 1450056, 2014.
[39] C. Sibona and S. Walczak, "Unfriending on Facebook: Friend request and online/offline behavior analysis," in *Proc. IEEE 44th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2011, pp. 1–10.
[40] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 2, p. 025102, 2001.
[41] L. A. Adamic and E. Adar, "Friends and neighbors on the Web," *Soc. Netw.*, vol. 25, no. 3, pp. 211–230, 2003.
[42] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, Oct. 2009.
[43] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
[44] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 4, p. 046122, 2009.
[45] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.
[46] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw.*, vol. 56, no. 18, pp. 3825–3833, 2012.
[47] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, Mar. 2017.
[48] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.
[49] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.



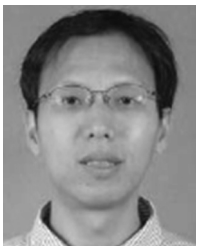
BO XU received the B.Sc. and Ph.D. degrees from the Dalian University of Technology, China, in 2007 and 2014, respectively. She is currently a Lecturer with the School of Software, Dalian University of Technology. Her current research interests include social computing, data mining, information retrieval, and natural language processing.

LU LI received the master's degree with the School of Software, Dalian University of Technology. She takes part in designing algorithms and conducting experiments in this paper. Her research interest is big data processing and analysis.

JIAYING LIU received the B.S. degree in software engineering from the Dalian University of Technology, China, in 2016, where she is currently pursuing the Ph.D. degree with the School of Software. Her research interests include big scholarly data, social network analysis, and science of success.



LIANGTIAN WAN (M'15) received the B.S. and Ph.D. degrees from the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China, in 2011 and 2015, respectively. From 2015 to 2017, he has been a Research Fellow with the School of Electrical and Electrical Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Software, Dalian University of Technology, China. He has published over 40 scientific papers in international journals and conferences. His research interests include social network analysis and mining, big data, array signal processing, wireless sensor networks, and compressive sensing and its application. He is an Associate Editor of the IEEE Access.



XIANGJIE KONG (M'13–SM'17) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor with the School of Software, Dalian University of Technology, China. He has published over 50 scientific papers in international journals and conferences (with over 30 indexed by ISI SCIE). His research interests include intelligent transportation systems, mobile computing, and cyber-physical systems. He is a Senior Member of CCF and a member of ACM. He has served as a (guest) editor of several international journals, the workshop chair, or a PC member of a number of conferences.

FENG XIA (M'07–SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently a Full Professor with the School of Software, Dalian University of Technology, China. He has published two books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, network science, data science, and mobile social networks. He is a Senior Member of ACM.